DOCUMENT RESUME

ED 077 934 TM 002 729

AUTHOR	Osborn, William C.
TITLE	Developing Performance Tests for Training Evaluation.
INSTITUTION	Human Resources Research Organization, Alexandria, Va.
SPONS AGENCY	Office of the Chief of Research and Development (Army), Washington, D.C.
REPORT NO	HumRRO-PP-3-73
PUB DATE	Feb 73
NOTE	8p.; Paper presented at U.S. Continental Army Command Training Workshop (Fort Gordon, Georgia, October 1971)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Military Personnel; *Military Training; Performance Specifications; *Performance Tests; Scoring; Speeches; *Student Evaluation; Task Performance; *Test Construction; Training Objectives

ABSTRACT

This paper describes the major action points in the course of developing a test for training evaluation. The author gives a brief summary of the 14 action points he considers basic for a test developer: (1) obtain list of terminal objectives with skill and knowledge requirements; (2) determine criticality of objectives to military mission; (3) determine adequacy of objective: presence of task behavior, conditions and standard; (4) review objective with job/training analyst; (5) determine feasibility of duplicating the objective's conditions and task behavior in a test situation; (6) develop a substitute method of testing: simulating conditions or task behavior; (7) determine number of replications or variations of test behavior necessary for reliable measurement; (8) determine controls on test conditions necessary to insure standardization over trainees; (9) develop objective pass-fail scoring procedure for trainee qualification; (10) develop diagnostic scoring procedures for training evaluation; (11) prepare detailed instructions for tester, trainee, and scorer; (12) determine feasibility of testing on all terminal objectives; (13) determine a relevant sample of test items (terminal objectives) for inclusion in test; and (14) prepare final specifications for test administration. (Author/KM)

Professional Paper 3-73
HumRRO-PP-3-73

HumRRO

Developing Performance Tests for Training Evaluation

William C. Osborn

Presentation at
U.S. Continental Army Command
Training Workshop
Fort Gordon, Georgia October 1971

HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street • Alexandria, Virginia 22314

February 1973

Prepared for

Office of the Chief of Research and Development
Department of the Army
Washington, D.C. 20310

The Human Resources Research Organization (HumRRO) is a nonprofit corporation established in 1969 to conduct research in the field of training and education. It is a continuation of The George Washington University Human Resources Research Office. HumRRO's general purpose is to improve human performance, particularly in organizational settings, through behavioral and social science research, development, and consultation. HumRRO's mission in work performed under contract with the Department of the Army is to conduct research in the fields of training, motivation and leadership.

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Published
February 1973
by
HUMAN RESOURCES RESEARCH ORGANIZATION
300 North Washington Street
Alexandria, Virginia 22314

Prefatory Note

This paper was presented at the U.S. Continental Army Command Training Workshop at Fort Gordon, Georgia in October 1971. The research on which this paper is based was performed under Work Unit TRAINMAN, Development of an Instructional Program in Training Technology and Training Management, at the Human Resources Research Organization, Division No. 2, Fort Knox, Kentucky.

DEVELOPING PERFORMANCE TESTS FOR TRAINING EVALUATION

William C. Osborn

A performance test is a template—a template modeled from a job task and used to gauge the similarity of a trained behavior to the demands of that job task. This view of performance tests implies a straightforward approach to their development. One simply re-creates the circumstances of the job task, asks the trainee to perform the task, and then records that he did or did not do it. Unfortunately, from our own experience we know that it is not this simple. Many practical problems intervene to complicate the process. We often find that a job has so many tasks that days would be needed to test them all. Occasionally, the equipment, terrain, and other support requirements prevent a realistic test for even a single task. At other times, we run into standards of task performance that are difficult to translate into a pass-fail criterion for scoring. We also have found that trainers need more than pass-fail results; they need diagnostic information to tell them why their trainees failed, if they did.

These are some of the major problems encountered by test developers, though by no means all. For the most part, the kinds of test development problems that we encounter in the field of training evaluation are not the same as those encountered in the field of aptitude testing. Thus, we have found the traditional body of academic literature on test development to be poorly suited to our needs. Certainly the basic notions of reliability and validity apply to any test development effort, but in our field, the exotic, sophisticated formulas that fill most books on test development are of little use.

One vital need in the field of training evaluation, it seems to me, is a how-to-do-it manual for test developers—one that responds to the variety of practical constraints and problems that occur in the process of constructing tests for the myriad tasks spanned by some eight or nine hundred Army jobs.

I wish that I had such a manual for you, but I don't. What I do have is intended to be a step, albeit small, in that direction. I have attempted to chart the major action points in the course of developing a test for training evaluation. These steps in performance test development are shown in Figure 1, and I hope that you will find it a useful framework for discussing the problems and practices of test development.

There are two matters of terminology that need clarification. The first has to do with the concept of performance testing. I choose to use this concept (at least today) to designate the test or tests, normally developed and administered by a quality control agency on completion of training for the two explicit purposes of qualifying trainees and evaluating training. This type of testing is to be distinguished from the development and use of tests by trainers for monitoring student progress within and between stages of training. The second is that I use the term *test item* in referring to the evaluation of behavior involved in a single job task, and the term *test* in referring to the aggregate of these items over an entire job or job sector purportedly covered by the training program. I am not asking you to agree with these labels, but to bear them in mind for the moment.

Now let us return to the process of test development as outlined in the figure. I should like to proceed through the 14 steps, and give a brief summary of my thoughts on the "why, what, and how" of each one.

Steps in Performance Test Development

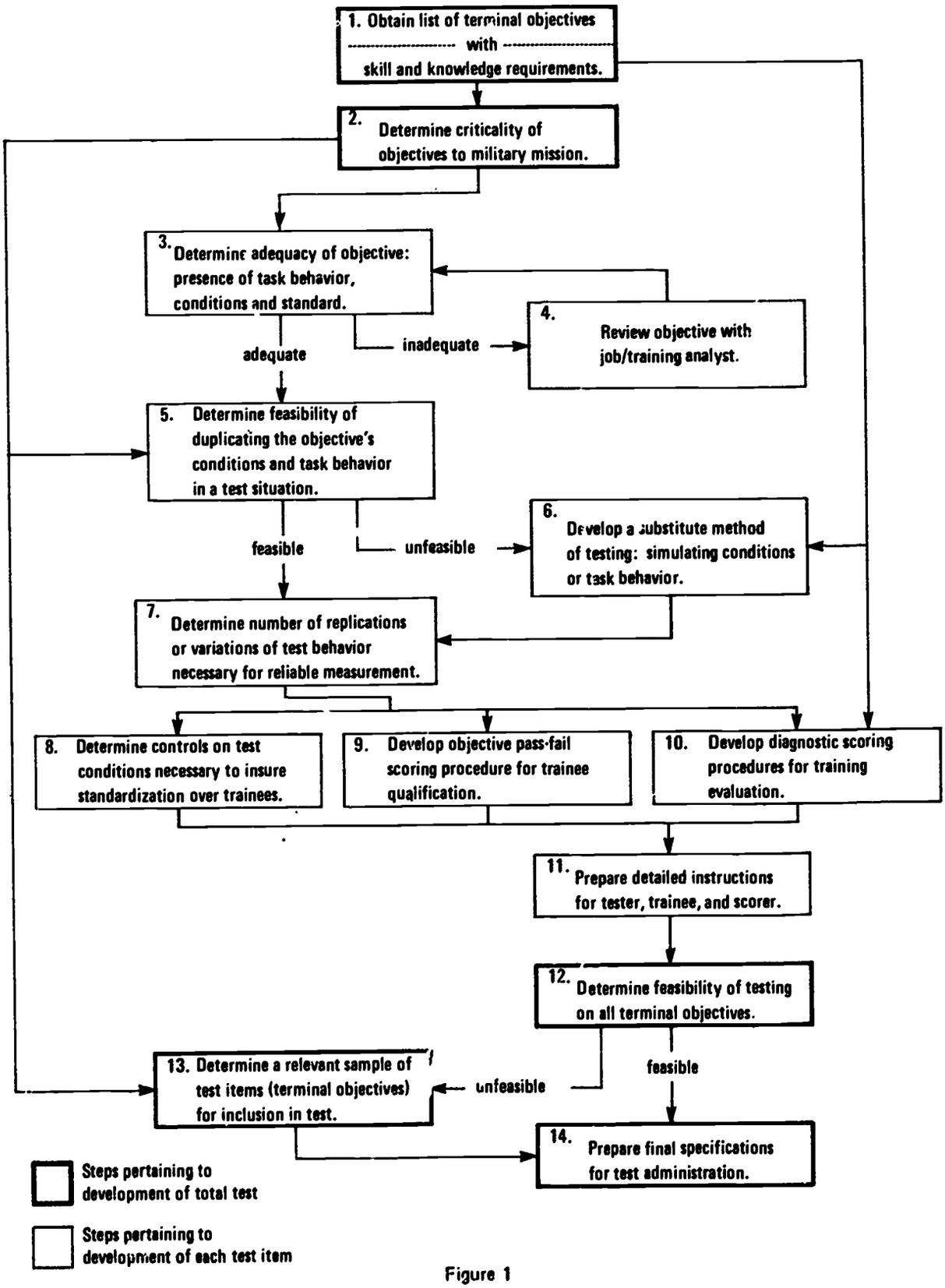

Figure 1

The first three steps on the chart concern assembling information that should routinely be supplied to the test developer. He should only have to verify completeness of the information, and not make judgments about its accuracy. As stated in the first step, test development begins with the objectives for the job or job sector for which people are to be trained. These are sometimes termed job objectives—more often, terminal training objectives. Whatever they are called, they are the master list of specifications derived from the job, and from which *both* training developers and performance test developers, sepatately, begin their work. As test developers, our goal is to develop a performance test item for each and every objective, although this is not to imply that our final test will necessarily encompass all objectives. In addition, each objective should be accompanied by a supporting list of skill and knowledge requirements to be used in later stages of test development.

The information designated in Step 2 should also be available as a matter of course. The relative importance of each objective, as judged in terms of mission capability, represents data that is necessary in making trade-offs later in the test development process.

Step 3 suggests that each objective must be reviewed to make sure it is all there. We know that, in addition to a stated task behavior, an objective should contain stated conditions and standards of performance. If any of the three elements are missing, or if any are unclear to the test developer, he should get together with the task analyst and, as indicated in Step 4, obtain a clear statement of the missing or confusing elements. Performance standards are the most common source of trouble, and if a fair and meaningful pass-fail criterion is to be established for a test item, the developer must have an unequivocal standard of task performance to work from.

In Step 5, test item development really begins. Here, the developer must judge the feasibility of duplicating in a test situation the conditions and behavior called for in the objective. Normally, of course, our view is that well stated objectives are blueprints for testing—in fact, dictating what the test conditions will be. Occasionally, however, we encounter an objective calling for the use of job-relevant equipment, terrain, support personnel, or a time frame that exceeds the resources available to the test agency. In these instances, the developer must carefully weigh the criticality of the objective (from Step 2) against the cost factors before deciding that full realism cannot be afforded, because invariably some degree of relevance is lost as one departs from the test specifications given in the objective.

When it is decided that the conditions of the objective cannot be duplicated in the test situation, a substitute technique must be developed, as indicated in Step 6. This is perhaps the most subtle and challenging aspect of the development process. Here, a developer's inventiveness is often needed in devising a method and conditons for testing that will call for the demonstration of a behavior that is as similar as possible to the behavior stated in the objective. Too often in this situation developers resort to paper-and-pencil tests measuring knowledge of the task, an approach that in most cases can be safely rejected out of hand. In considering simulation options developers have a useful check available in the task's skill and knowledge requirements. The relevance of a proposed test method may be evaluated by checking the number of skill and knowledge components of the task that are called for in the method.

Once a task-relevant method of testing is determined, Step 5 or Step 6, the developer turns his attention to the matter of achieving measurement reliability. In Step 7, he must again look at the objective in terms of repetitions or variations of the behavior implied. In most cases, this will be explicitly given. For a specific skill, such as disassembling a rifle or installing a carburetor, a single demonstration of the behavior is all that is normally called for. On occasion, however, with generalized skills or generalized behaviors, the number of repetitions of the behavior may or may not be clearly stated in

the objective. An objective specifying that something will be done correctly 9 out of 10 times creates no problem for the test item developer, as 10 repetitions are required. On the other hand, the standard may be phrased in terms of correct performance on 90% of the trials. Here a decision must be reached on an appropriate number of repetitions of the performance to ask for in the test item. More generally, the important consideration in Step 7 is whether a large enough sample of trainee performance is being required so that success or failure does not result largely from chance. Here, again, the test developer must make some trade-off between time or cost factors and reliability of the measured behavior.

Step 8 pertains to another aspect of test reliability—the standardization of the conditions under which a test item is administered. Here, the important factors are the instructions and environmental conditions under which the test item is given. Instructions should be identical for everyone. They should be clearly and simply stated, leaving nothing to the interpretation or misinterpretation of the trainees taking the test. Things such as the method of scoring and whether speed or accuracy is important should be stressed in the instructions. Also, conditions pertaining to test supplies and environmental factors should be constant for all personnel. Items of equipment worked with or on during testing should be restored to their pretest condition if they are used by successive trainees. Similarly, environmental factors such as visibility, temperature, attitude of the tester, time of day, and the like, must be stabilized.

In Step 9, a final aspect of measurement reliability is considered. Here procedures for translating an observed trainee performance into a pass-fail score must be developed. Provision for this type of scoring should be structured so that only the more reliable human skills are used. That is, the scoring activity should be reduced to one of matching or comparing the test item response with some model of the acceptable response. If the model response on a test of rifle marksmanship is defined as a hole in the bullseye, then the scorer has a relatively easy task in judging the acceptability of the response made by the rifleman. Unfortunately, responses for many test items cannot be judged in this "either/or" fashion, but require a "more-or-less" type of judgment. In these cases, the developer should always strive to break down the model response into elements so that comparative judgments can be made more easily by the scorer. This may often entail preparing a checklist of the necessary components or features of the model response.

In Step 10, a supplementary scoring procedure is developed for use in diagnosing reasons for trainee failure on the test item. Pass-fail scoring is sufficient in meeting the primary mission of quality control, which is the certification of trainee job readiness. However, the secondary mission, that of training program evaluation, is best accomplished by providing the trainers not only with the incidence of pass and failure for an objective, but also feedback on why trainees failed. One way to obtain this data is through a checklist developed from the skill and knowledge requirements of the task to be used by the tester in recording why the trainee failed a test item. When accumulated over a number of test item administrations, this diagnostic information will normally provide a stable picture of the reasons for failure that trainers may then use to selectively revise and strengthen their program.

In Step 11, the test developer simply brings together the products of previous steps and formats the final test item. Detailed instructions to the tester covering test materials, equipment, procedures, precautions, and so forth, are spelled out. The directions to be read to the trainee by the tester, and the scoring procedure should also be written out.

The final three steps in the figure pertain to assembly and administration of the final form of the test. In Step 12, a decision is made on whether time permits testing on all objectives—that is, administration of all test items. If it is not feasible to do so, an appropriate sample of test items has to be selected (Step 13). As indicated in this step, the main criterion for sampling should derive from criticality ratings of the objectives. An

exact procedure for doing this will depend upon the categories originally used for reporting criticality. Generally, the developer would first include all "essential" or highly critical items, and then sample from the remaining. Wherever sampling is necessary, the usual practice is to vary the sample from one administration to the next so that all test items are used sooner or later. Variations in the sample should not be systematic in the sense that trainers or trainees can anticipate what items are going to appear.

In Step 14, final guidance for test administration is prepared. Training for testers may have to be developed; lists of equipment and materials prepared; and scheduling worked out. If testing is to be done individually, it is usually a good idea to prescribe a "county fair" layout of test stations. This serves purposes of economy, as well as permitting test items to be administered in varying order. In addition, security precautions must be specified to ensure, for example, that one trainee cannot benefit by observing another's performance, or that trainees do not talk among themselves during test administration.

Consideration of these action points, step by step, constitutes a framework for performance test development.

Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Human Resources Research Organizaton (HumRRO) 300 North Washington Street Alexandria, Virginia 22314	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

DEVELOPING PERFORMANCE TESTS FOR TRAINING EVALUATION

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Professional Paper

5. AUTHOR(S) *(First name, middle initial, last name)*

William C. Osborn

6. REPORT DATE February 1973	7a. TOTAL NO. OF PAGES 8	7b. NO. OF REFS 0
8a. CONTRACT OR GRANT NO. DAHC-19-73-C-0004 b. PROJECT NO. 2Q062107A745 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) HumRRO-PP-3-73 9b. OTHER REPORT NO.(S) *(Any other numbers that may be assigned this report)*	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES Presented at the CONARC Training Workshop Fort Gordon, Georgia. HumRRO Division No. 2, Fort Knox, Kentucky	12 SPONSORING MILITARY ACTIVITY Office, Chief of Research and Development Department of the Army Washington, D.C. 20310

13. ABSTRACT

This paper describes the major action points in the course of developing a test for training evaluation. The author gives a brief summary of the 14 action points he considers basic for a test developer, from job objectives to final specifications.

DD FORM 1 NOV 68 1473

Unclassified

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Learning systems						
Performance tests						
Training evaluatio-						